⑤ 層別法について

青山　博次郎

§1　緒　言

標本抽出に際して精度をあげるため層別が行われるのが普通である。　このときどのような規準をえらべばよいか，またその規準について実際上どのように層別を行えば最も精度があがるかについて考察してみよう。

§2.　層別規準のえらび方

我々が問題とする標識を x とし，これと相関の高い標識が以前の調査または資料によって知られているものとし，その二つを仮りに y, z としよう。　このとき如何なる y, z をとることによって層別の効果が上がるであろうか。　これを重相関係数によって測ることにしてみよう。

y, z に対する x の重相関係数を R とし，x, y の相関係数，x, z の相関係数，y, z の相関係数をそれぞれ r_{12}, r_{13}, r_{23} とする。　このとき

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2 r_{12} r_{13} r_{23}}{1 - r_{23}^2} \qquad (1)$$

が成立する。　便宜上 $f = R^2$ とおいて f の最大値が如何なる r_{12}, r_{13}, r_{23} によって得られるかを調べるのである。

さて

$$r_{12} = \rho \cos\theta, \qquad r_{13} = \rho \sin\theta \tag{2}$$

とおくと

$$f = \rho^2 \frac{1 - r_{23} \sin 2\theta}{1 - r_{23}^2}$$

$$= \rho^2 \frac{1 - k r_{23}}{1 - r_{23}^2} \tag{3}$$

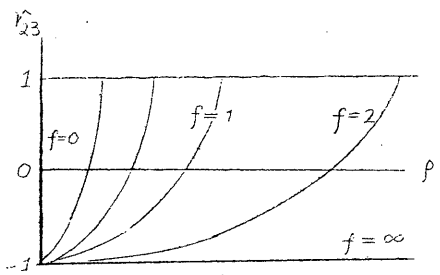
但し　　$k = \sin 2\theta$

(3)或いは(1)は通常の微分法による極値の求め方では極小値，または極大でも極小でもない値しか求められないので直接的に考えてみる。

(i) $k = 1$ のとき

$$f = \frac{\rho^2}{1 + r_{23}}$$

となるから $f = const.$ の等高線は放物線をえがく。



(ii) $k = -1$ のとき

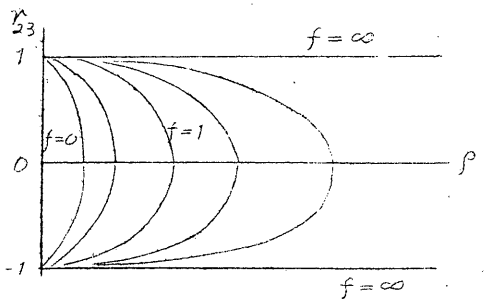
$$f = \frac{\rho^2}{1 - r_{23}}$$

となるから同様な上下反対の放物線をえがく。

(iii) $k = 0$ のとき

$$f = \frac{\rho^2}{1 - r_{23}^2}$$

となるから楕円群が等高線を與える。



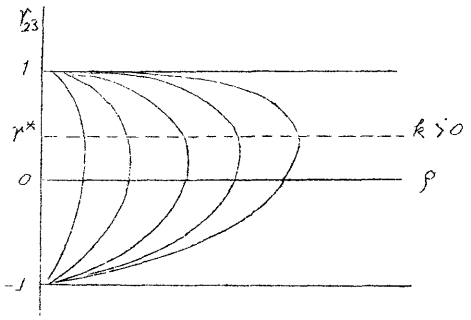
(iV) 一般の k については

$$f = \rho^2 \frac{1 - k r_{23}}{1 - r_{23}^2}$$

であるから k の正，負によってそれぞれ $r_{23} = 1, -1$ よりに ρ の最大値をとる曲線群が得られる。

従って $k = \pm 1$ 即ち $r_{12} = \pm r_{13}$ なるときは $r_{23} = \mp 1$ のとき f が最大となり，然らざる場合は $r_{23} = \pm 1$ に近い程 f は大きくなる。

例えば下表のようになつている。



$$r^* = \frac{1}{k} - \sqrt{\frac{1}{k^2} - 1}$$

r_{12}	r_{13}	r_{23}	$f = R^2$	R	k
.4	.4	.9	.1684	.41	
		.4	.2286	.48	
		.0	.3200	.57	1
		−.4	.5333	.72	
		−.9	3.2000	1.79	
.4	.2	.9	.2947	.54	
		.4	.1619	.40	
		.0	.2000	.45	≠0, ±1
		−.4	.3143	.55	
		−.9	1.8105	1.35	
.4	.0	±.9	.8421	.92	0
		±.4	.1905	.44	
		± 0	.1600	.40	

実際には $r_{12}=r_{13}=.4$ で $r_{23}=-.9$ というようなことは起り得ないから $0 \leqq R \leqq 1$ の範囲内でのみ考えればよい。

これらの結果より層別の基準とすべき標識 y, z が各々 x に対して同程度の相関をもつならば y, z 相互の相関は小さい程よく，y, z の各々が x に対する相関に差のあるときは，y, z の相互の相関は大きい程よいことが分る。

規準とすべき標識が2つ以上になるときは更に複雑な様相を示すであろうが，2つの場合の考え方を逐次拡張して行けばよいであろう。

例えば y, z なる2つの層別の規準が既に得られたとき第三の規準 u を加えて x に対する重相関係数を更に大きくすればよい。u についての相関係数には添字4を用いると，重相関係数の平方の増加量 f は

$$f=\frac{\left(r_{14}(1-r_{23}^2)-r_{24}(r_{12}-r_{13}r_{23})-r_{34}(r_{13}-r_{12}r_{23})\right)^2}{1-r_{23}^2-r_{24}^2-r_{34}^2+2r_{23}r_{24}r_{34}}$$

となる。このとき $f=Const.$ なる面は通常の場合 r_{14}, r_{24}, r_{34} を変数とする三次元空間の楕円面を表わすから r_{14} が大きいだけでなく，r_{14}, r_{24}, r_{34} の種々の組合せに対して f は最大となることが分る。

この操作は規準が更に増すときも同様であって，$n-1$ 個の規準に加えて更に一つ規準をますときの重相関係数の平方の増加量を f とおけば，

$$f=\frac{\left\{\alpha r_{1n}+L(r_{2n}, r_{3n}, \cdots, r_{n-1,n})\right\}^2}{\beta+Q(r_{2n}, r_{3n}, \cdots, r_{n-1,n})}$$

但し α, β は $r_{1n}, r_{2n}, \cdots, r_{n-1,n}$ についての定数，L, Q はそれぞれ一次式，二次式を表わす。

従って $f=Const.$ は超楕円面となり前と同様種々の組合せが考えられる。

§ 3. 一つの量 x を推定するための最適層別法。

推定すべき量 x と相関の高い量 y について母集団を層別することを考えよう。 y の分布函数 $F(y)$ は分っていて，これを L 個の層に別け，各層より比例抽出法を用いて $unit$ を抽出し，そのとき推定総量 x の抽出分散を最小にするような層別法を求めるのが目的である。 考え方は林・丸山両氏の論文[註1)]と同様である。

ただ直接 x の資料は層別に用いられない点を考慮したのである。

x の分散は一般に

$$V(x) = \sum_{i=1}^{L} N_i^2 \frac{N_i - n_i}{N_i - 1} \frac{\sigma_{ix}^2}{n_i} \tag{4}$$

但し母集団の総数を N，第 i 層の数を N_i，第 i 層よりのサンプル数を n_i，第 i 層における x の母分散を σ_{ix}^2 とする。

比例抽出法を用いるときは

$$n_i = n \frac{N_i}{N} \tag{5}$$

であるから，有限母集団修正を無視すれば (4) は

$$V(x) = \frac{N}{n} \sum_i N_i \sigma_{ix}^2 \tag{6}$$

いま x と y の間に線形相関 $x = \alpha' y + \beta'$ が成立つならば，近似的に定数 α を用いて

$$\sigma_x^2 = \alpha^2 \sigma_y^2 \tag{7}$$

が成立つと考えてよい。 従ってまた第 i 層に於ても

$$\sigma_{ix}^2 = \alpha^2 \sigma_{iy}^2 \tag{7}'$$

が成立つと考えても差支えはないであろう。

さて $F(y)$ は密度函数を有するとし，第 i 層での平均を $\bar{Y}_i$，全分散を σ_y^2，全平均を $\bar{Y}$ とし，各層の分点を y_i とおくと

註 1) 林・丸山：ある層化法に就て，統数研彙究録，第4巻第10号，1948

$$\left.\begin{aligned} \sigma_y^2 &= \int_{-\infty}^{\infty} (y-\bar{Y})^2 \, d\Phi(y) \\ \bar{Y} &= \int_{-\infty}^{\infty} y \, d\Phi(y) \\ \bar{Y}_i &= \int_{y_{i-1}}^{y_i} y \, d\Phi(y) \Big/ (\Phi(y_i) - \Phi(y_{i-1})) \end{aligned}\right\} \qquad (8)$$

このとき

$$N\sigma_y^2 = \sum_i N_i \sigma_{iy}^2 + \sum_i N_i (\bar{Y} - \bar{Y}_i)^2 \qquad (9)$$

を用いると (6), (7) より

$$V(x) = \frac{N\alpha^2}{n} \sum_i N_i \sigma_{iy}^2 = \frac{N\alpha^2}{n} \left\{ N\sigma_y^2 - \sum_i N_i (\bar{Y} - \bar{Y}_i)^2 \right\}$$

$$= \frac{N\alpha^2}{n} \left\{ N\sigma_y^2 + N\bar{Y}^2 - \sum_i N_i \bar{Y}_i^2 \right\} \qquad (10)$$

となるから，$V(x)$ を最小にするには

$$f = \sum_{i=1}^{L} N_i \bar{Y}_i^2 = N \sum_i \left(\int_{y_{i-1}}^{y_i} y \, d\Phi \right)^2 \Big/ (\Phi(y_i) - \Phi(y_{i-1})) \qquad (11)$$

を最大にすればよい。 ここで $y_0 = -\infty$, $y_L = +\infty$ とする

L が一定のときは $\frac{\partial f}{\partial y_i} = 0$ より

$$y_i = \frac{\bar{Y}_i + \bar{Y}_{i+1}}{2}, \qquad i = 1, 2, \cdots, L-1 \qquad (12)$$

が得られる。

層の数を増せば (10) が小さくなることは (9) の関係式より

$$N\sigma^2 \geq N_1 \sigma_1^2 + N_2 \sigma_2^2 \quad (N = N_1 + N_2)$$

を使うことによって容易に分る。

また (12) による分点の位置は逐次近似的に求めるより方法はないが，それらの相対的位置については次のことがいえる。

(12) を書き直せば

$$2y_i = \frac{\int_{y_{i-1}}^{y_i} y\,d\Phi}{\Phi(y_i)-\Phi(y_{i-1})} + \frac{\int_{y_i}^{y_{i+1}} y\,d\Phi}{\Phi(y_{i+1})-\Phi(y_i)}$$

平均値定理を用いるとき

$$(y_i - y_{i-1})(1-\theta) = (y_{i+1}-y_i)\theta' \qquad (13)$$

但し　$0<\theta<1,\quad 0<\theta'<1$

或いは近似的に

$$(y_i - y_{i-1})\frac{\Phi'(y_{i-1})}{2\Phi'(y_{i-1}^*)} = (y_{i+1}-y_i)\left(1-\frac{\Phi'(y_i)}{2\Phi'(y_i^*)}\right) \qquad (14)$$

但し　$y_{i-1} < y_{i-1}^* < y_i < y_i^* < y_{i+1}$

それ故第 i 層で y の密度函数 $\Phi'(y)$ の変化が小なるとき

$$y_i - y_{i-1} \fallingdotseq y_{i+1} - y_i$$

即ち等間隔となる。　この関係はまた層の幅が小さくなるときに於ても成立する。

§4.　二量を規準とする最適層別法

推定すべき総量を Z とし，これと相関の高い二量を x, y として，これらによる最適層別法を考える。　前と同様にサンプルの割当は比例法を用いるとすれば

$$V(\bar{z}) = \frac{N}{n}\sum_{i=1}^{L} N_i \sigma_{iz}^2 \qquad (15)$$

となる。　また線形相関　$z = \alpha' x + \beta' y + \gamma'$ が成立つものと仮定すれば，前と同様の記号を用いて（ρ_{ixy} は第 i 層内の x と y の相関係数）

$$\sigma_{iz}^2 = \alpha^2\sigma_{ix}^2 + \beta^2\sigma_{iy}^2 + 2\alpha\beta\rho_{ixy}\sigma_{ix}\sigma_{iy} \qquad (16)$$

このとき

$$\left.\begin{aligned} N\sigma_x^2 &= \sum_i N_i \sigma_{ix}^2 + \sum_i N_i (\bar{x} - \bar{x}_i)^2 \\ N\sigma_y^2 &= \sum_i N_i \sigma_{iy}^2 + \sum_i N_i (\bar{y} - \bar{y}_i)^2 \\ \sum_i N_i \rho_{ixy} \sigma_{ix} \sigma_{iy} &= \sum_i N_i \bar{x}_i \bar{y}_i + N\rho_{xy}\sigma_x\sigma_y - N\bar{x}\bar{y} \end{aligned}\right\} \quad (17)$$

従って（15）は（16），(17) により

$$V(\bar{z}) = \frac{N}{n}\left\{\alpha^2(N\sigma_x^2 + N\bar{x}^2 - \sum_i N_i \bar{x}_i^2) + \beta^2(N\sigma_y^2 + N\bar{y}^2 - \sum_i N_i \bar{y}_i^2) + 2\alpha\beta(N\rho_{xy}\sigma_x\sigma_y - N\bar{x}\bar{y} + \sum_i N_i \bar{x}_i \bar{y}_i)\right\} \quad (18)$$

故に $V(\bar{z})$ を最小にするためには（L を一定として）

$$f = \alpha^2 \sum_i N_i \bar{x}_i^2 + \beta^2 \sum_i N_i \bar{y}_i^2 - 2\alpha\beta \sum_i N_i \bar{x}_i \bar{y}_i \quad (19)$$

を最大にすればよい。

ここで前節の如く x, y の分布函数をそれぞれ $\Phi(x)$, $\Psi(y)$ とし，何れも密度函数をもつとすれば

$$\frac{\partial f}{\partial x_i} = 0, \qquad \frac{\partial f}{\partial y_i} = 0 \quad \text{より}$$

$$x_i = \frac{\alpha(\bar{x}_i^2 - \bar{x}_{i+1}^2)}{2\{\alpha(\bar{x}_i - \bar{x}_{i+1}) - \beta(\bar{y}_i - \bar{y}_{i+1})\}} \quad (20)$$

$$y_i = \frac{\beta(\bar{y}_i^2 - \bar{y}_{i+1}^2)}{2\{\beta(\bar{y}_i - \bar{y}_{i+1}) - \alpha(\bar{x}_i - \bar{x}_{i+1})\}} \quad (21)$$

が成立する。　両式より α, β を消去すれば

$$y_i = -\frac{\bar{y}_i + \bar{y}_{i+1}}{\bar{x}_i + \bar{x}_{i+1}}\left(x_i - \frac{\bar{x}_i + \bar{x}_{i+1}}{2}\right) \quad (22)$$

或いは

$$x_i = -\frac{\bar{X}_i+\bar{X}_{i+1}}{\bar{Y}_i+\bar{Y}_{i+1}}\left(y_i-\frac{\bar{Y}_i+\bar{Y}_{i+1}}{2}\right) \qquad (23)$$

(22)または(23)式より各層の境界点 (x_i, y_i) は一直線上にあり，各軸とは前節とは同様の点

$$x_i = \frac{\bar{X}_i+\bar{X}_{i+1}}{2} \quad 及び \quad y_i = \frac{\bar{Y}_i+\bar{Y}_{i+1}}{2}$$

に於いて交わる。

故に各層はこれらの直線群で境された帯状領域と考えることが出来る。

尚規準とすべき量が三つ以上の場合も全く同様な考え方を用いることが出来る。

§5 Neyman 法を用いる場合

前節まですべて標本割当は比例法であったが，Neyman 法を用いた時を参考のためにあげておこう。推定すべき量は x，層別の基準とする量は y とすると，推定総量の分散は

$$V(x) \fallingdotseq \frac{1}{n}\left(\sum_i N_i \sigma_{ix}\right)^2 \qquad (24)$$

前と同様にして x と y との線形相関を仮定すれば

$$V(x) = \frac{\alpha^2}{n}\left(\sum_i N_i \sigma_{iy}\right)^2 \qquad (25)$$

従って

$$f = \sum_i N_i \sigma_{iy} = N\sum_i \sqrt{(\Phi(y_i)-\Phi(y_{i-1}))\int_{y_{i-1}}^{y_i}(y-\bar{Y}_i)^2 d\Phi} \qquad (26)$$

を最小になる如く y_i を求めればよい。

$\frac{\partial f}{\partial y_i} = 0$ より

$$\sigma_{iy}\left\{1+\left(\frac{y_i-\bar{Y}_i}{\sigma_{iy}}\right)^2\right\} = \sigma_{i+1,y}\left\{1+\left(\frac{y_i-\bar{Y}_{i+1}}{\sigma_{i+1,y}}\right)^2\right\} \qquad (27)$$

が成立する如く y_i をえらべばよい。　このときも逐次近似法によらねばならない。(註2)

平均値定理を用いると (27) は,

$$(\theta''-\theta)(y_i-y_{i-1})\left(1+\left(\frac{1-\theta\cdot}{\theta''-\theta}\right)^2\right)=(\theta'''-\theta')(y_{i+1}-y_i)\left(1+\left(\frac{\theta'}{\theta'''-\theta'}\right)^2\right) \tag{28}$$

但し　$0<\theta,\theta',\theta'',\theta'''<1$

このときも各層内で $\overline{\Phi}'(y)$ の変化が小なるときは

$$y_i-y_{i-1}\fallingdotseq y_{i+1}-y_i$$

即ち等間隔となる。

また層の幅が小さくなるときは $\theta,\theta'\to\frac{1}{2}$　$\theta''-\theta,\ \theta'''-\theta'\to\sqrt{3}/6$ となることが証明せられるので，同様に等間隔となる。

§6.　一般の場合

前節と同様にして x の総量を推定するのに y を層別の規準にとると

$$V(x)\fallingdotseq\sum_i N_i^2\frac{\sigma_{ix}^2}{n_i}=\alpha^2\sum_i N_i^2\frac{\sigma_{iy}^2}{n_i}$$

$$=\alpha^2N^2\sum_i\frac{\overline{\Phi}(y_i)-\overline{\Phi}(y_{i-1})}{n_i}\int_{y_{i-1}}^{y_i}(y-\overline{Y}_i)^2d\overline{\Phi} \tag{29}$$

従って $\frac{\partial V}{\partial y_i}=0$　より

$$\frac{N_i}{n_i}(\sigma_{iy}^2+(y_i-\overline{Y}_i)^2)=\frac{N_{i+1}}{n_{i+1}}(\sigma_{i+1,y}^2+(y_i-\overline{Y}_{i+1})^2) \tag{30}$$

これより y について既に層別を施してあるときは σ_{iy}^2, $\overline{Y}_i$ 等を計算し逐次 (30) により n_i を定めればよい。

比例抽出法，Neyman 法を用いるときも (30) より分割点 y_i

(註) 2) T. Dalenius : The Problem of Optimum Stratification, Skandinavisk Aktuarietidskrift, 1950

を求めると前と同様の結果が得られる